SKRINING THALASSEMIA BERBASIS DATA HEMATOLOGIS MENGGUNAKAN MODEL RANDOM FOREST DAN SUPPORT VECTOR MACHINE

Mifta Mardiah, Niken Satuti Nur Handayani, Azhari, Rarastoeti Pratiwi

INTISARI

Thalassemia merupakan kelainan genetik akibat gangguan produksi rantai globin penyusun hemoglobin dan menjadi tantangan kesehatan masyarakat di seluruh dunia karena dampaknya terhadap peningkatan angka kejadian dan kematian. Pemeriksaan thalassemia terdiri dari tiga level, pemeriksaan hingga Level III dilakukan apabila data hematologis tidak memberikan petunjuk status thalassemia. Kehadiran *Artificial Intelligence* (AI)—khususnya model *Random Forest* dan *Support Vector Machine*—menawarkan alternatif untuk mempercepat skrining thalassemia menggunakan data hematologis (Level I). Penelitian ini bertujuan menerapkan model AI dalam analisis data hematologis untuk skrining thalassemia. Data hematologis diperoleh dari Grup Riset Thalassemia Universitas Gadjah Mada. Data *preprocessing* dilakukan sebelum *clustering* dan klasifikasi. *K-means* dan *Gaussian Mixture Model* (GMM) digunakan sebagai metode *clustering* untuk menemukan pola dan mendeteksi anomali data. Pelatihan uji model menggunakan *Random Forest* dan *Support Vector Machine*. Model dievaluasi melalui hasil *confusion matrix*. Hasil *clustering* menggunakan *k-means clustering* membentuk 6 *cluster* sesuai jumlah kategori, dengan data normal dan β-thalassemia terpisah cukup jelas. Sebagian data β-thalassemia dan HbE, serta kategori lain menunjukkan tumpang tindih yang mengindikasikan kedekatan karakteristik hematologi. *Clustering* GMM mencapai akurasi 83.72%. Hasil pemodelan menunjukkan *Random Forest* memperoleh akurasi tertinggi sebesar 97%, sedangkan pada model SVM, kernel RBF menghasilkan akurasi tertinggi sebesar 87%. Di sisi lain, SVM-Linear memberikan performa terbaik (55,81%). Selain itu, hanya *Random Forest* dan SVM-Linear yang mampu mengenali kategori β-thalassemia, menunjukkan perbedaan sensitivitas antar-model. Hasil ini menunjukkan bahwa integrasi *machine learning* dapat membantu efisiensi skrining dengan mempercepat alur pemeriksaan thalassemia yang umumnya membutuhkan pemeriksaan lanjutan.

Kata Kunci : Thalassemia, *Random Forest*, *Support Vector Machine*, Data Hematologis, Skrining

THALASSEMIA SCREENING BASED ON HEMATOLOGICAL DATA USING RANDOM FOREST AND SUPPORT VECTOR MACHINE MODELS

Mifta Mardiah, Niken Satuti Nur Handayani, Azhari, Rarastoeti Pratiwi

ABSTRACT

Thalassemia is a genetic disorder caused by impaired production of globin chains that form hemoglobin and remains a global public health challenge due to its impact on morbidity and mortality. Thalassemia screening consists of three levels, in which examination up to Level III is required when hematological data do not provide any indication of thalassemia status. The presence of Artificial Intelligence (AI)—particularly Random Forest and Support Vector Machine—offers an alternative approach to accelerate thalassemia screening using hematological data (Level I). This study aims to apply AI models in analyzing hematological data for thalassemia screening. Data preprocessing was performed before clustering and classification. K-means and Gaussian Mixture Model (GMM) were used as clustering methods to find pattern and detect data anomalies. Model training and testing used Random Forest and Support Vector Machine. The model was evaluated through the result of confusion matrix. The clustering result using k-means produced six clusters according to the number of categories, with normal and β-thalassemia data clearly separated. Several subsets of β-thalassemia, HbE, and other categories exhibited overlap, indicating similarities in hematological characteristics. GMM clustering achieved an accuracy of 83.72%. In the classification models, Random Forest achieved the highest accuracy of 97%, while the SVM model, the RBF kernel produced the highest accuracy of 87%. On the other hand, SVM-Linear provided the best performance (55,81%). Furthermore, only Random Forest and SVM-Linear were able to recognize the β-thalassemia category, indicating differences in sensitivity across models. These results indicate that machine learning integration can improve screening efficiency by accelerating the thalassemia examination process, which generally requires further examination.

Keywords : Thalassemia, Random Forest, Support Vector Machine, Hematological Data, Screening